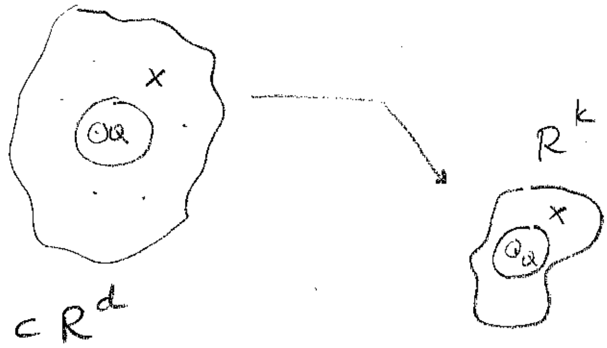
14.02.2020

Johnson - Lindenstrauss Theorem -

$k << d$

$d >> 1$



$A \subset R^d$

$R^d \ni a \xrightarrow{\pi} R^k$

Isometry: L_2-norm

Given a $\varepsilon \in (0,1)$ a map $f: R^d \to R^k$ is called an

ε- isometry if $\forall a, a' \in A$.

$$(1-\varepsilon)\|a-a'\|^2 \leq \|f(a)-f(a')\|^2 \leq (1+\varepsilon)\|a-a'\|^2$$

$$l(c_1 a + c_2 b) = c_1 \cdot l(a) + c_2 l(b)$$

$$\pi(v_1+v_2) = \pi(v_1) + \pi(v_2)$$

* Which k?
  - choose vectors randomly.
  - Concentration estimates: that my points mainly concentrate around the mean.

    If I assign a gaussian measure on my entire space and

    choose

    — I take a big space, in each coordinate I assign gaussian measure

    * unit variance) gaussian measure
    * zero mean

    — each.

iid. X_{ij} $\quad i=1,\dots,k$ and $j=1,\dots,d$.

$E[X_{ij}]=0 \qquad \mathrm{var}(X_{ij})=1$

→ definition

Vector $v \in \mathbb{R}^d$, $\qquad \Pi_i(v) \equiv \sum_{j=1}^{d} v_j x_{ij} \qquad i=1,\dots,k$

$= (v_1, v_2, \dots, v_d)$

$$\Pi(v) = (\Pi_1(v), \dots, \Pi_k(v))$$

$$E\left[\|\Pi(v)\|^2\right] = E\left[\sum_{i=1}^{k} \|\Pi_i(v)\|^2\right] = E\left[\sum_{i=1}^{k}\sum_{j=1}^{d} v_j^2 x_{ij}^2\right] = k\|v\|^2$$

$$= \sum_{i=1}^{k} \|\Pi_i(v)\|^2 = \sum_{i=1}^{k}\sum_{j=1}^{d} v_j^2 \underbrace{x_{ij}^2}_{\text{physical coordinates ?}} + \underbrace{2\sum_{i=1}^{k}\sum_{\substack{l,m=1 \\ l\neq m}}^{d} v_l v_m x_{il} x_{im}}_{=0}$$

\- Distribution -

which gives the concentration around the mean ⇒ Gaussian.

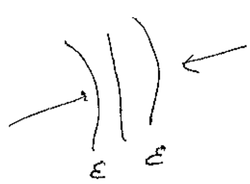
$$\Pi_i(v) = \frac{1}{\sqrt{k}} \sum_{j=1}^{d} v_j x_{ij} \qquad i=1,\dots k$$

$v_1 - v_2$

$$E\|\Pi(v)\|^2 = \frac{1}{k} E\left[\sum_{i=1}^{k} \|\Pi_i(v)\|^2\right]$$

$v_1 - v_2$

$$= \frac{1}{k} E\left[\sum_{i=1}^{k} \|\Pi_i(v)\|^2\right] = \frac{1}{k} E\left[\sum_{i=1}^{k}\sum_{j=1}^{d} v_j^2 x_{ij}^2\right] = \|v\|^2$$



?

Gaussian (Centered) — Same machinery can be used.

$M_X(t) = e^{t^2/2}$, Sub-gaussian $M_X(t) \leq e^{t^2/2}$

* Gaussian Annulus Theorem:

For a d-dim'l. spherical Gaussian with unit variance in each direction for $\beta \leq \sqrt{d}$, all but at most $3e^{x\beta^2}$ of the prob. mass is in the annulus $\sqrt{d} - \beta \leq \|x\| \leq \sqrt{d} + \beta$, c is a fixed true number

$\gamma = \|X\|$ $\quad \sqrt{d} - \beta \leq \gamma \leq \sqrt{d} + \beta$ $\quad |\gamma - \sqrt{d}| \geq \beta$

$|\gamma^2 - d| \geq \beta(\gamma + \sqrt{d}) \geq \beta\sqrt{d}$

$Y_i = X_i^2 - 1$, $\gamma^2 - d = Y_1 + Y_2 + \dots + Y_d$.

$|Y_1 + Y_2 + \dots + Y_d| \geq \beta\sqrt{d}$ $\quad E[Y_i] = 0$

- For $|X_i| \leq 1$ $\quad |Y_i|^s \leq 1$ $\Rightarrow$ $|E[Y_i^s]| = E[|Y_i|^s] \leq E[1 + X_i^{2s}]$

$|X_i| \geq 1$ $\quad |Y_i|^s \leq |X_i|^{2s}$, $= 1 + \sqrt{\frac{2}{\pi}} \int_0^\infty x^{2s} e^{-x^2/2} dx$

$E[|Y_i|^s] \leq 2^s \cdot s!$

$E(Y_i) = 0$

$Var(Y_i) \leq 2^2 . 2! = 8$

$W_i = \frac{Y_i}{2}$, $Var(W_i) \leq 2$, $E[W_i^s] \leq 2 s!$

$$|w_1, w_2, \dots w_d| \geq \frac{\beta\sqrt{d}}{2}$$

$Var(w) = \sigma^2 = 2$ $\quad n = d$

$3e^{-\frac{\beta^2 d}{|x|^2 2}} \longrightarrow ?$ $\quad |\sqrt{d}|$, $E[|X|^2] = d$.

* The order tail estimates will not change

*

$\Pi : \mathbb{R}^d \to \mathbb{R}^k$ $\quad u_1, \dots, u_k \in \mathbb{R}^d$

$\Pi(v) = (u_i v, \dots, u_k v)$ $\quad u_i = (X_{i1}, X_{i2}, \dots X_{id})$

Theorem: (Random Projection Theorem) (R.P. Theorem)

Fix $v \in \mathbb{R}^d$. Define Π as above $\exists c > 0$ s.t

$\forall \varepsilon \in (0,1)$, $P[|\,|\Pi(v)| - \sqrt{k}\,|v|\,| \geq \varepsilon\sqrt{k}\,|v|] \leq 3e^{-ck\varepsilon^2}$

– $\Pi(v)$ is a random spherical Gaussian with unit variance $k = d$.

(?) Johnson – Lindenstrauss Theorem:

For $0 < \varepsilon < 1$ and any integer $d >> 1$ with $k \gg \frac{3}{c\varepsilon^2} \ln N$.

N is the number of points

for $\Pi : \mathbb{R}^d \longrightarrow \mathbb{R}^k$ wp $1 - \frac{3}{2N}$

$$(1-\varepsilon)\sqrt{k}\,|v_i - v_j| \leq |\Pi(v_i) - \Pi(v_j)| \leq (1+\varepsilon)\sqrt{k}\,|v_i - v_j| \quad \forall v_i, v_j \in \mathbb{R}^d$$

$\llcorner$ = distance between query and data is smaller than ε in lower-dim'l space.

$$P\left[\,|\Pi(v_i - v_j)| \notin \left[(1-\varepsilon)\sqrt{k}\,|v_i - v_j|,\ (1+\varepsilon)\sqrt{k}\,|v_i - v_j|\right]\right] \leq 3\underbrace{e^{-ck\varepsilon^2}}$$

$\underbrace{\binom{N}{2}}$ $O(N^2) \longrightarrow \underbrace{\frac{3}{2} N^2 \cdot \frac{1}{N^3}} = \frac{3}{2N}$

we have N-points

$\binom{N}{2}$ pairs

$$\frac{3}{N^3} \cdot \underbrace{\left(\frac{N^2}{2}\right)}_{\sim \binom{N}{2}}$$

$$e^{-c\varepsilon^2 \cdot \frac{3}{c\varepsilon^2} \ln N} = e^{\ln N^{-3}} = \frac{1}{N^3}$$

- query & data lie in any dimension. (even in ∞-dim.)
- ∞-dimension version of Euclidean norm. = Hilbert's space
- Square-integrable function. $\rightarrow$?